

JP200142310

Publication Title:

DEVICE AND METHOD FOR RECOGNIZING CHARACTER STRING, AND
COMPUTER PROGRAM

Abstract:

Abstract of JP 2002245408

(A) Translate this text PROBLEM TO BE SOLVED: To perform character recognition of even an image data obtained from a recording medium having stains, etc., with high accuracy. SOLUTION: In a candidate character string dictionary file XDb1, character strings to be candidates for character recognition are previously stored as candidate character string data. Character string extraction image data PDL are extracted from a scanner image obtained by performing optical reading and are image data for a recognition object. The image data PDL are equally divided into the number of parts which coincides with the number of characters of character string pattern data 'Katakura' in Japanese in a direction (horizontal) in which the character string lines. Individual character extraction image data PDC obtained by equaling dividing the image data PDL are respectively subjected to pattern collation with a standard character pattern specified by character data having the same arrangement sequences as those of candidate character string pattern data.

Courtesy of <http://v3.espacenet.com>

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2002-245408
(P2002-245408A)

(43) 公開日 平成14年8月30日 (2002.8.30)

(51) Int.Cl. ⁷	識別記号	F I	デマコト* (参考)		
G 0 6 K	9/72	C 0 6 K	9/72	Λ	5 B 0 2 9
	9/34		9/34		5 B 0 6 4
	9/62	6 2 0	9/62	6 2 0 D	

審査請求 有 請求項の数18 O L (全 19 頁)

(21) 出願番号 特願2001-42310(P2001-42310)

(22) 出願日 平成13年2月19日 (2001.2.19)

(71) 出願人 000002369

セイコーエプソン株式会社

東京都新宿区西新宿2丁目4番1号

(72) 発明者 中島 雄二

長野県松本市中央二丁目1番27号 エー・

アイ ソフト株式会社内

(74) 代理人 100096817

弁理士 五十嵐 孝雄 (外3名)

Fターム(参考) 5B029 AA01 BB02 CC21 EE08

5B064 AA01 BA01 CA08 DA03 DA14

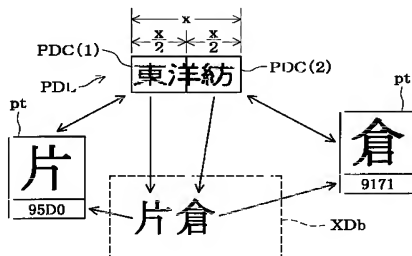
EA19 EA36

(54) 【発明の名称】 文字列認識装置およびその方法並びにコンピュータプログラム

(57) 【要約】

【課題】 汚れ等がある記録媒体から得られた画像データであっても高精度の文字認識を可能とする。

【解決手段】 候補文字列辞書ファイルXD b 1 には、文字認識の候補となる文字列が候補文字列データとして予め記憶されている。文字列抽出画像データPDLは、光学的に読み取って得られたスキャナ画像から抽出したもので認識対象の画像データである。文字列抽出画像データPDLをその文字列の並び方向（横方向）に、文字列パターンデータ「片倉」の文字数と一致する数に等分割する。そして、等分割して得られた各文字抽出画像データPDCを、候補文字列パターンデータの同じ配列順位の文字データにより特定される標準文字パターンとそれぞれパターン照合する。



【特許請求の範囲】

【請求項１】 光学的に読み取って得られた画像データを文字列データに変換する文字列認識装置であって、１文字毎の標準文字パターンを記憶する標準文字パターン記憶部と、

認識結果として出力されるべき文字列を候補文字列データとして記憶する候補文字列記憶手段と、

前記画像データの中から、前記候補文字列記憶手段に記憶された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す部分画像データ切り出し手段と、

前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう照合手段と、

前記照合手段の照合結果に基づいて、前記候補文字列データを交換結果として出力する出力制御手段とを備えることを特徴とする文字列認識装置。

【請求項２】 前記照合手段は、前記切り出した部分画像データ毎に、前記候補文字列データの内の当該部分画像データと同じ配列順位の文字データとの間で前記照合を行なう構成である請求項１に記載の文字列認識装置。

【請求項３】 請求項１または２に記載の文字列認識装置であって、前記候補文字列記憶手段は、前記候補文字列データを複数記憶する構成であり、

前記候補文字列記憶手段に記憶された複数の候補文字列データ毎に、前記部分画像データ切り出し手段および照合手段および出力制御手段をそれぞれ動作させる手段を備える文字列認識装置。

【請求項４】 請求項３に記載の文字列認識装置であって、

前記候補文字列記憶手段は、同一の範疇に含まれる複数の前記候補文字列データから構成される候補文字列群を複数組記憶する構成であり、

前記複数組の候補文字列群から一組を選択する候補文字列群選択手段を備えるとともに、

前記選択された候補文字列群の候補文字列データを、前記部分画像データ切り出し手段、照合手段および出力制御手段に利用する構成とした文字列認識装置。

【請求項５】 新たな候補文字列群を前記候補文字列記憶手段に追加する手段を備える請求項４に記載の文字列認識装置。

【請求項６】 請求項１ないし５のいずれかに記載の文字列認識装置であって、

前記部分画像データ切り出し手段により切り出した部分画像データの切り出し幅を微調整して新たな部分画像データを生成する手段と、

前記生成された新たな部分画像データを利用して、前記照合手段および出力制御手段を再動作させる手段とを備える文字列認識装置。

【請求項７】 請求項１ないし６のいずれかに記載の文字列認識装置であって、

前記画像データの中から、空白部分に基づいて１文字に相当する部分画像データを切り出す第１手段と、

前記第１手段により切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう第２手段と、

前記第２手段の照合結果が認識不可である場合に、前記部分画像データ切り出し手段および照合手段および出力制御手段をそれぞれ動作させる第３手段とを備える文字列認識装置。

【請求項８】 光学的に読み取って得られた画像データを記憶手段に格納される１文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する文字列認識方法であって、（ａ）認識結果として出力されるべき文字列を候補文字列データとして予め記憶するステップと、（ｂ）前記画像データの中から、前記ステップ（ａ）により記憶された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出すステップと、（ｃ）前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なうステップと、（ｄ）前記ステップ（ｃ）による照合結果に基づいて、前記候補文字列データを交換結果として出力するステップとを備えることを特徴とする文字列認識方法。

【請求項９】 請求項８に記載の文字列認識方法であって、（ｅ）前記画像データの中から、空白部分に基づいて１文字に相当する部分画像データを切り出すステップと、（ｆ）前記ステップ（ｅ）により切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なうステップと、（ｇ）前記ステップ（ｆ）による照合結果が認識不可である場合に、前記ステップ（ｂ）ないし（ｄ）をそれぞれ実行させるステップとを備える文字列認識方法。

【請求項１０】 光学的に読み取って得られた画像データを記憶手段に格納される１文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する処理を実行するコンピュータプログラムであって、（ａ）認識結果として出力されるべき文字列を候補文字列データとして記憶手段から読み出す機能と、（ｂ）前記画像データの中から、前記機能（ａ）により読み出された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す機能と、（ｃ）前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、（ｄ）前記機能（ｃ）による照合結果に基づいて、前記候補文字列データを変換結果として出力する機能とを、コンピュータに実現させるためのコンピュータプログラム。

【請求項１１】 前記機能（ｃ）は、前記切り出した部

分画像データ毎に、前記候補文字列データの内の当該部分画像データと同じ配列順位の文字データとの間で前記照合を行なうものである請求項10に記載のコンピュータプログラム。

【請求項12】 請求項10または11に記載のコンピュータプログラムであって、(e) 候補文字列データを複数備える記憶手段から候補文字列データが読み出されるように前記機能(a)を繰り返し実行させる機能と、(f) 前記機能(e)によって読み出された候補文字列データ毎に、前記機能(b)ないし(d)をそれぞれ繰り返し実行させる機能とを、コンピュータに実現させるためのコンピュータプログラム。

【請求項13】 請求項12に記載のコンピュータプログラムであって、(g) 前記記憶手段に記憶された、同一の範疇に含まれる複数の前記候補文字列データから構成される複数組の候補文字列群から一組を選択する機能を、コンピュータに実現させるとともに、前記選択された候補文字列群の候補文字列データを、前記機能(b)ないし(d)の実現に利用する構成としたコンピュータプログラム。

【請求項14】 新たな候補文字列群を記憶手段に追加する機能をさらにコンピュータに実現させるための請求項13に記載のコンピュータプログラム。

【請求項15】 請求項10ないし14のいずれかに記載の文字列認識装置であって、(h) 前記機能(b)により切り出した部分画像データの切り出し幅を微調整して新たな部分画像データを生成する機能と、(i) 前記生成された新たな部分画像データを利用して、前記機能(c)および機能(d)を再動作させる機能とを、コンピュータに実現させるためのコンピュータプログラム。

【請求項16】 請求項10ないし15のいずれかに記載のコンピュータプログラムであって、(j) 前記画像データの中から、空白部分に基づいて1文字に相当する部分画像データを切り出す機能と、(k) 前記機能

(j)により切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、(l) 前記機能(k)による照合結果が認識不可である場合に、前記機能(b)ないし(d)をそれぞれ実行させる機能とを、コンピュータに実現させるためのコンピュータプログラム。

【請求項17】 光学的に読み取って得られた画像データを記憶手段に格納される1文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する処理を実行するコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体であって、(a) 認識結果として出力されるべき文字列を候補文字列データとして記憶手段から読み出す機能と、(b) 前記画像データの中から、前記機能(a)により読み出された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す機能と、(c) 前記切り出した部分画

像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、(d) 前記機能(c)による照合結果に基づいて、前記候補文字列データを変換結果として出力する機能とを、コンピュータに実現させるためのコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項18】 請求項17に記載のコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体であって、(e) 前記画像データの中から、空白部分に基づいて1文字に相当する部分画像データを切り出す機能と、(f) 前記機能(e)により切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、(g) 前記機能(f)による照合結果が認識不可である場合に、前記機能(b)ないし(d)をそれぞれ実行させる機能とを、コンピュータに実現させるためのコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、文字列を光学的に読み取って得られた画像データを文字列データに変換する文字列認識の技術に関する。

【0002】

【従来の技術】従来より、文字認識の技術として、新聞、雑誌等の一般文書の記録紙をイメージスキャナ等によって光学的に読み取って、その得られた記録紙の画像データを文字列データに変換する技術が知られている。この技術は、上記記録紙の画像データから1文字に相当する部分画像データを切り出して、これらを予め用意した候補文字パターンと照合することで類似度が高い候補文字パターンを選び出して、その選び出した候補文字パターンを認識結果として出力するものである。

【0003】

【発明が解決しようとする課題】しかしながら、上記従来の技術では、光学的に読み取って得られた画像データが、記録された文字が細かい記録紙からのものである場合、記録紙に汚れ等があると、隣り合う2文字を1文字に相当すると誤判定して文字認識の精度が極端に低下するという問題があった。

【0004】この発明は、上記問題に鑑みてなされたもので、汚れ等がある記録紙から得られた画像データであっても高精度の文字認識を可能とすることを目的としている。

【0005】

【課題を解決するための手段およびその作用・効果】前述した課題の少なくとも一部を解決するための手段として、以下に示す構成をとった。

【0006】この発明の文字列認識装置は、光学的に読み取って得られた画像データを文字列データに変換する

文字列認識装置であって、1文字毎の標準文字パターンを記憶する標準文字パターン記憶部と、認識結果として出力されるべき文字列を候補文字列データとして記憶する候補文字列記憶手段と、前記画像データの中から、前記候補文字列記憶手段に記憶された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す部分画像データ切り出し手段と、前記切り出した部分画像データ毎に、前記候補文字列データのうちより特定される前記標準文字パターンとの照合を行なう照合手段と、前記照合手段の照合結果に基づいて、前記候補文字列データを変換結果として出力する出力制御手段とを備えることを特徴としている。

【0007】上記構成（以下、基本構成と呼ぶ）の文字列認識装置によれば、部分画像データ切り出し手段によって、画像データ全体に対する各部分画像データの占有割合を、候補文字列データにおける1文字の占める割合と一致させて部分画像データの切り出しを行なうことができる。このために、画像が記録された記録紙に汚れ等があっても隣り合う2文字分を1文字に相当すると誤判定することもない。したがって、文字認識の精度が極端に低下することを防止することができる。

【0008】上記基本構成の文字認識装置において、前記照合手段は、前記切り出した部分画像データ毎に、前記候補文字列データの内の当該部分画像データと同じ配列順位の文字データとの間で前記照合を行なう構成とすることができる。

【0009】この構成によれば、一の部分画像データに対して照合させる文字パターンの数を減らすことができることから、認識速度を向上することができる。

【0010】上記基本構成の文字認識装置において、前記候補文字列記憶手段は、前記候補文字列データを複数記憶する構成であり、前記候補文字列記憶手段に記憶された複数の候補文字列データ毎に、前記部分画像データ切り出し手段および照合手段および出力制御手段をそれぞれ動作させる手段を備える構成とすることもできる。

【0011】この構成によれば、複数の候補文字列データの中から照合結果の優れた候補文字列データを選んで、その選んだ候補文字列データを変換結果として出力することが可能となる。

【0012】候補文字列データを複数記憶する上記構成の文字列認識装置において、前記候補文字列記憶手段は、同一の範疇に含まれる複数の前記候補文字列データから構成される候補文字列群を複数組記憶する構成であり、前記複数組の候補文字列群から一組を選択する候補文字列群選択手段を備えるとともに、前記選択された候補文字列群の候補文字列データを、前記部分画像データ切り出し手段、照合手段および出力制御手段に利用する構成とすることができる。

【0013】この構成によれば、候補文字列群選択手段によって一組の候補文字列群を選択することで、その選

択された候補文字列群によって示される範疇の文字列を高精度に認識することができる。

【0014】上記候補文字列群選択手段によって一組の候補文字列群を選択可能とした文字列認識装置において、新たな候補文字列群を前記候補文字列記憶手段に追加する手段を備える構成とすることができる。この構成によれば、認識結果として出力されるべき文字列を自在に追加することができる。

【0015】上記基本構成の文字列認識装置において、前記部分画像データ切り出し手段により切り出した部分画像データの切り出し幅を微調整して新たな部分画像データを生成する手段と、前記生成された新たな部分画像データを利用して、前記照合手段および出力制御手段を再動作させる手段とを備える構成とすることができる。この構成によれば、画像データにて示される文字のフォントがプロポーションナル文字である場合にも、高精度に文字を認識することができる。

【0016】上記基本構成の文字列認識装置において、前記画像データの中から、空白部分に基づいて1文字に相当する部分画像データを切り出す第1手段と、前記第1手段により切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう第2手段と、前記第2手段の照合結果が認識不可である場合に、前記部分画像データ切り出し手段および照合手段および出力制御手段をそれぞれ動作させる第3手段とを備える構成とすることができる。

【0017】この構成によれば、上記基本構成による文字認識の前に、画像データの中から空白部分に基づいて切り出した部分画像データ毎に、候補文字列データにより特定される標準文字パターンとの照合を行なう、この照合結果が認識不可となった場合に、上記基本構成による文字認識が行なわれる。したがって、画像データの中から空白部分に基づいて部分画像データを切り出す一般的な切り出し方法に基づく文字認識と、上記基本構成による文字認識との2段階をもって文字認識を行なうことから、認識精度を一層高めることができる。

【0018】この発明の文字列認識方法は、光学的に読み取って得られた画像データを記憶手段に格納される1文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する文字列認識方法であって、

(a) 認識結果として出力されるべき文字列を候補文字列データとして予め記憶するステップと、(b) 前記画像データの中から、前記ステップ(a)により記憶された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出すステップと、(c) 前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なうステップと、(d) 前記ステップ(c)による照合結果に基づいて、前記候補文字列データを変換結果として出力するステップとを備えることを特徴としている。

【0019】上記構成の文字列認識方法は、上記発明の文字列認識装置と同様な作用・効果を有しており、汚れ等がある記録紙から得られた画像データを認識する場合であっても、文字認識の精度が極端に低下することを防止することができる。

【0020】この発明のコンピュータプログラムは、光学的に読み取って得られた画像データを記憶手段に格納される1文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する処理を実行するコンピュータプログラムであって、(a)認識結果として出力されるべき文字列を候補文字列データとして記憶手段から読み出す機能と、(b)前記画像データの中から、前記機能(a)により読み出された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す機能と、(c)前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、(d)前記機能(c)による照合結果に基づいて、前記候補文字列データを変換結果として出力する機能とを、コンピュータに実現させることを特徴としている。

【0021】この発明の記録媒体は、光学的に読み取って得られた画像データを記憶手段に格納される1文字毎の標準文字パターンと照合して、該画像データを文字列データに変換する処理を実行するコンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体であって、(a)認識結果として出力されるべき文字列を候補文字列データとして記憶手段から読み出す機能と、(b)前記画像データの中から、前記機能(a)により読み出された候補文字列データの文字数と一致する数に等分割された部分画像データを切り出す機能と、(c)前記切り出した部分画像データ毎に、前記候補文字列データにより特定される前記標準文字パターンとの照合を行なう機能と、(d)前記機能(c)による照合結果に基づいて、前記候補文字列データを変換結果として出力する機能とを、コンピュータに実現させるためのコンピュータプログラムを記録したことを特徴としている。

【0022】上記構成のコンピュータプログラムおよび記録媒体も、上記発明の文字列認識装置や文字列認識方法と同様な作用・効果を有しており、汚れ等がある記録紙から得られた画像データを認識する場合であっても、文字認識の精度が極端に低下することを防止することができる。

【0023】

【発明の他の態様】この発明は、以下のような他の態様も含んでいる。その第1の態様は、この発明のコンピュータプログラムを含むことで搬送波内に具現化されたデータ信号としての態様である。第2の態様は、コンピュータプログラムを通信経路を介して供給するプログラム供給装置としての態様である。この第2の態様では、コンピュータプログラムをネットワーク上のサーバなどに

置き、通信経路を介して、必要なプログラムをコンピュータにダウンロードし、これを実行することで、上記の装置や方法を実現することができる。

【0024】

【発明の実施の形態】以上説明したこの発明の構成・作用を一層明らかにするために、以下この発明の実施の形態を実施例に基づき説明する。

【0025】1. ハードウェアの全体構成

図1は、この発明の一実施例を適用するコンピュータシステムのハードウェアの概略構成を示すブロック図である。このコンピュータシステムは、いわゆるパーソナルコンピュータ(以下、単にコンピュータと呼ぶ)を中心に備え、その周辺にCRTディスプレイ12およびイメージスキャナ14を備える。コンピュータは、コンピュータ本体16とキーボード18とマウス20を備える。なお、このコンピュータ本体16には、CD-ROM22の内容を読み取るCDドライブ24が搭載されている。

【0026】コンピュータ本体16は、中央演算処理装置としてのCPU30を中心にバスにより相互に接続されたROM31、RAM32、表示画像メモリ33、マウスインタフェース34、キーボードインタフェース35、CDC36、HDC37、CRTC38、入出力機用インタフェース40およびI/Oポート41を備える。ROM31は、内蔵されている各種プログラム等を記憶する読み出し専用のメモリである。RAM32は、各種データ等を記憶する読み出し・書込み可能なメモリである。表示画像メモリ33は、CRTディスプレイ12に表示する画像の画像データを記憶するメモリである。マウスインタフェース34は、マウス20とのデータ等のやり取りを司るインタフェースである。キーボードインタフェース35は、キーボード18からのキー入力力を司るインタフェースである。CDC36は、CDドライブ(CDD)24を制御するCDコントローラである。HDC37は、ハードディスクドライブ(HDD)42を制御するハードディスクコントローラである。

【0027】CRTC38は、表示画像メモリ33に記憶される表示画像データに基づいてCRTディスプレイ12における画像の表示を制御するCRTコントローラである。入出力機用インタフェース40は、外部に接続された入出力機器、この実施例ではイメージスキャナ14へのデータの入出力を制御するインタフェースである。I/Oポート41は、シリアル出力のポートを備えており、モデム44に接続されており、このモデム44を介して、公衆電話回線46に接続されている。コンピュータ本体16は、モデム44を介して、外部のネットワークに接続されておき、特定のサーバ47に接続可能となっている。

【0028】このコンピュータシステムでは、オペレーティングシステムはHDD42に記憶されており、コン

ビュータ本体 16 に電源を投入すると、HDD42 のブートブロックに書き込まれたローグに従って RAM32 の所定の領域にロードされる。また、イメージスキャナ 14 で取り込んだ画像（以下、スキャン画像と呼ぶ）をテキストデータに変換する文字認識用ソフトウェア（コンピュータプログラム）は、CD-ROM22 に予め格納されており、所定のインストールプログラムを起動することによって、CDドライブ 24 からビュータ本体 16 にインストールされる。このインストールされたコンピュータプログラムは、HDD42 に記憶されており、所定の起動命令を受けたときに、RAM32 の所定の領域にロードされる。

【0029】このコンピュータプログラムを CPU30 が実行することによって本発明の各種構成要件は実現される。このコンピュータプログラムは、前述したように、CD-ROM22 に格納されたものであるが、これに替えて、フロッピーディスク、光磁気ディスク、IC カード等の他の携帯型記録媒体（可搬型記録媒体）に格納された構成としてもよい。また、前述したコンピュータプログラムは、外部のネットワークに接続される特定のサーバ 47 から、ネットワークを介して提供されるプログラムデータをダウンロードして、RAM32 または HDD42 に転送することにより得るようにすることもできる。なお、上記ネットワークとしては、インターネットであってもよく、特定のホームページからダウンロードして得たコンピュータプログラムであってもよい。あるいは、電子メールの添付ファイルの形態で供給されたコンピュータプログラムであってもよい。

【0030】以上説明したハードウェア構成を有するコンピュータシステムによる文字認識の様子について次に説明する。図 2 は、コンピュータ本体 16 によって実行される文字認識処理の様子を示すブロック図である。図示するように、この文字認識処理に用いられるデータファイルを記憶する構成として、HDD42 には、候補文字列記憶部 48 と標準文字パターン記憶部 49 が備えられる。

【0031】候補文字列記憶部 48 は、認識結果として出力されるべき文字列を候補文字列として記憶するもので、候補文字列を示すテキストデータの集合である候補文字列辞書が、予め複数種類記憶されている。一般に、OCR（光学式文字読取装置）においては、漢字（JIS 第 1 水準 + JIS 第 2 水準）、ひらがな、カタカナ、アルファベット、数字等、一般に使用される全ての文字を認識文字として、これら認識文字の全ての組合せを候補文字列として網羅するが、この文字認識用ソフトウェア 50 は、特定の限定された集合に含まれる文字列だけを認識結果とすることで、認識精度の向上を図っている。ここで、特定の限定された集合としては、予め定めた集まりであらばどのようなものでもよく、例えば、都道府県名の集合、証券銘柄の集合等が該当する。

【0032】標準文字パターン記憶部 49 には、上述した一般に使用される全ての文字の標準文字パターンがビットマップデータの形で予め記憶されている。標準文字パターン記憶部 49 に格納される各標準文字パターンには、対応するテキストデータが予め記されており、標準文字パターン記憶部 49 にテキストデータを投入することにより、そのテキストデータにより特定される標準文字パターンを読み出すことが可能となっている。

【0033】コンピュータ本体 16 の内部で動作している文字認識用ソフトウェア 50 によれば、まず、スキャン画像取込部 51 によりスキャナドライブ 60 を動作させてイメージスキャナ 14 から新聞、雑誌などの画像（スキャン画像）を取り込む処理を行なう。

【0034】次いで、スキャン画像取込部 51 によって取り込まれたスキャン画像から文字認識の対象となる任意の画像領域を、領域抽出部 52 によって抽出する。また、候補文字列指定部 53 によって、候補文字列記憶部 48 に格納される候補文字列辞書を複数種類の中から一つ選択することにより、認識結果として出力されるべき候補文字列群を指定する。

【0035】続いて、候補文字列読出部 54 によって、候補文字列記憶部 48 内の前記候補文字列指定部 53 で指定された候補文字列辞書の中から候補文字列を順に読み出す。第 1 認識部 55 は、領域抽出部 52 によって抽出された領域の画像データについての文字認識を行なう第 1 の認識処理を実行する。この第 1 の認識処理では、候補文字列読出部 54 によって読み出された候補文字列により特定される前記標準文字パターンとの照合を行なうことにより文字認識を行なう。第 1 認識部 55 で文字認識に成功した場合には、認識結果である候補文字列（テキストデータ）を出力する。一方、第 1 認識部 55 で文字認識に失敗した場合には、第 2 認識部 57 により、領域抽出部 52 によって抽出された領域の画像データについての文字認識を行なう第 2 の認識処理を行なう。この第 2 の認識処理は、第 1 認識部 55 による第 1 の認識処理とは文字認識のアルゴリズムが相違するものである。なお、この第 2 の認識処理でも、候補文字列読出部 54 によって読み出された候補文字列により特定される前記標準文字パターンとの照合を行なうことにより文字認識を行なう。

【0036】第 2 認識部 57 で文字認識に成功した場合には、認識結果である候補文字列（テキストデータ）を出力する。一方、第 2 認識部 57 で文字認識に失敗した場合には、エラーである旨のデータを出力する。文字列出力部 56 は、第 1 認識部 55 および第 2 認識部 57 から出力されるテキストデータを、ワープロソフトウェア 70 に送る。ワープロソフトウェア 70 は、ディスプレイドライバ 80 を介して CRT ディスプレイ 12 へ上記テキストデータを表示する。なお、この実施例では、送り先はワープロソフトウェア 70 としたが、これに替

て表計算ソフトウェア等の他のソフトウェアとすることもできる。

【0037】コンピュータ本体16のCPU30で上記文字認識用ソフトウェア50を実行することで、上述した文字認識を実現している。上記文字認識用ソフトウェア50に従う制御処理について、以下詳細に説明する。図3は、この制御処理のルーチンを示すフローチャートである。このルーチンは、文字認識用ソフトウェア50を実行させる旨の指示がなされた以後、所定時間毎に繰り返し実行される。

【0038】図示するように、処理が開始されると、CPU30は、まず、この文字認識用ソフトウェア50の起動後、最初であるか否かを判別する（ステップS100）。ここで、最初であると判別されたときには、初期画面を示すウィンドウW0をCRTディスプレイ12に表示する処理を行なう（ステップS110）。図4は、このウィンドウW0を示す説明図である。図示するように、初期画面のウィンドウW0には、メニューバーBR1とツールバーBR2が設けられている。ツールバーBR2には、作業手順に従った順にコマンドを実行するための「スキャン」、「領域抽出」、「認識」、「保存」のボタンBT1、BT2、BT3、BT4が設けられている。なお、図3に戻って、ステップS100で最初でないとは判別されたときには、ステップS110の処理は実行しない。

【0039】次いで、CPU30は、ツールバーBR2上の「スキャン」のボタンBT1が、マウス20によってクリック操作されたか否かを判別する（ステップS120）。ここで、肯定判別されたときには、CPU30は、文字認識の対象となる新聞、雑誌などの原稿PPをスキャンするスキャン処理を実行する（ステップS130）。

【0040】図5は、原稿PPの一例を示す説明図である。図5に例示するように、原稿PPは、新聞の株価掲載欄であり、例えば、東京第1部の株価が掲載されている。ステップS130では、CPU30は、スキャナドライバ60を動作させて、この原稿PPがセットされたイメージスキャナ14からこの原稿PPの画像を示すスキャン画像データを取り込む。ステップS130の実行後、「リターン」に抜けてこの処理を一旦終了する。

【0041】一方、ステップS120で否定判別されたときには、CPU30は、ステップS140に処理を進めて、「領域抽出」のボタンBT2が、マウス20によってクリック操作されたか否かを判別する。ここで、肯定判別されたときには、CPU30は、後述する領域抽出処理のルーチンを実行する（ステップS150）。一方、ステップS140で否定判別された場合には、ステップS160に処理を進める。なお、フローチャートには詳細に記載されていないが、ステップS140で肯定判別された場合でも、ステップS130のスキャン処理

の実行後でない場合には、ステップS160に処理を進める。

【0042】図6は、ステップS150で実行される領域抽出処理ルーチンの詳細を示すフローチャートである。図示するように、この領域抽出処理ルーチンに処理が移行すると、CPU30は、まず、ステップS130のスキャン処理により取り込まれたスキャン画像データから、文字列認識の対象となる任意の画像領域を抽出する処理を行なう（ステップS200）。作業者は、ウィンドウW0に表示された原稿PPのスキャン画像に対して、マウス20を用いて、文字列認識の対象とした画像領域の範囲を指定する操作を行なう。CPU30は、この作業によるマウス操作を受けて、そのマウス20によって指定された画像領域のデータをスキャン画像データの中から抽出する。一例として、作業者は、CRTディスプレイ12の画面上でマウス20を用いて、図7に示すように、原稿PPのスキャン画像の中から、東京第1部の「繊維」の銘柄を表わす範囲の画像領域PAを指定する操作を行なう。この場合、ステップS200では、その指定された画像領域PAの画像データが抽出される。この抽出した画像データは、抽出画像データPDとしてRAM32に格納される。

【0043】図6に戻り、ステップS200の実行後、CPU30は、抽出画像データをCRTディスプレイ12に表示する処理を行なう（ステップS210）。図8は、抽出画像データPDが表示されるウィンドウW1の一例を示す説明図である。図示するように、ウィンドウW1の左半分の領域WAに、ステップS200で抽出した抽出画像データPDが表示される。

【0044】図6に戻り、ステップS210の実行後、CPU30は、候補文字列辞書を複数の中から指定する処理を行なう（ステップS220）。候補文字列辞書は、抽出画像データの文字認識の処理を行なう際に使用する辞書であり、このコンピュータのHDD42に予め複数用意されている。ステップS220では、それら複数の候補文字列辞書の中から、ステップS200で抽出された抽出画像データPDの内容に対応する一つの候補文字列辞書が指定される。具体的には、その指定は、前述したウィンドウW1（図8）を用いた作業によるマウス操作に従って行なわれる。図8に示すように、ウィンドウW1の右欄には、候補文字列辞書選択用の入力欄IAが設けられており、その入力欄IAには、HDD42に格納されている候補文字列辞書の名前、例えば、「都道府県」、「繊維（東京第1部）」が表示される。作業者は、マウス20を用いたクリック操作により、入力欄IAから一の候補文字列辞書、すなわち、図8の領域WAに例示するように東京第1部の繊維の銘柄が抽出画像データPDとして表示されている場合には、入力欄IAにおいて「繊維（東京第1部）」を選択する。

【0045】なお、このウィンドウW1の右欄には、抽

出画像データPDに示される文字列が縦書きであるか横書きであるかを示すラジオボタンRBT1、RBT2も設けられており、ステップS220では、このラジオボタンRBT1、RBT2を用いて文字列が横書きであるか縦書きであるかの指定もなされる。ステップS220の実行後、この領域抽出処理ルーチンを一旦終了する。

【0046】図3に戻り、領域抽出処理ルーチンの実行後、「リターン」に抜けてこの処理を一旦終了する。ここまでで「領域抽出」の作業を終えたことになる。この時点でのHDD42およびRAM32に格納されるこの発明に関わるデータファイルを図9に示した。図9に示すように、HDD42には、「都道府県」の候補文字列辞書のファイルXDαと、「繊維（東京第1部）」の候補文字列辞書のファイルXDβと、標準文字パターンファイルPTFとがこの文字認識用ソフトウェア50で起動される前から格納されている。

【0047】「都道府県」の候補文字列辞書ファイルXDαには、都道府県名の候補文字列データがテキストデータの形で順に格納される。すなわち、候補文字列辞書ファイルXDαには、「北海道」、「青森県」、…といったテキストデータが順に格納される（図中、「北海道」、「青森県」といった文字で記したが、実際はテキストデータである文字コードが格納される）。「繊維（東京第1部）」の候補文字列辞書ファイルXDβには、繊維銘柄名の候補文字列データが順に格納される。すなわち、「片倉」、「グンゼ」、…といったテキストデータが順に格納される（図中、「片倉」、「グンゼ」といった文字で記したが、実際はテキストデータである文字コードが格納される）。標準文字パターンファイルPTFには、漢字（JIS第1水準+JIS第2水準）、ひらがな、カタカナ、アルファベット、数字等、一般に使用される全ての文字の標準文字パターンPTがビットマップデータの形で予め記憶されている。各標準文字パターンp_mには、対応するテキストデータt_xが予め対に記憶されており、テキストデータにより特定される標準文字パターンを読み出すことが可能となっている。

【0048】一方、RAM32には、ステップS200で得られた抽出画像データPDが格納される。この抽出画像データPDは、文字列認識を行なう対象としての画像データである。なお、上記HDD42には、「都道府県」と「繊維（東京第1部）」の候補文字列辞書ファイルXDα、XDβとが予め格納されていることを示したが、これら候補文字列辞書ファイルは別の処理ルーチンにより任意に追加できる構成とすることができる。

【0049】図10は、その別の処理ルーチンにより表示される追加用のウィンドウW2を示す説明図である。作業者は、このウィンドウW2上に設けられた「追加」のボタンBT11をクリックすることにより、新たな範疇の候補文字列辞書ファイルを登録することができる。

この構成によれば、認識結果として出力されるべき文字列を自在に追加することができる。

【0050】図3に戻り、ステップS160に処理が移行すると、ツールバーBR2上の「認識」のボタンBT3が、マウス20によってクリック操作されたか否かを判別する。ここで、肯定判別されたときには、CPU30は、抽出画像データPDをテキストデータに変換する後述する文字列認識処理のルーチンを実行する（ステップS170）。一方、ステップS160で否定判別された場合には、ステップS180に処理を進める。なお、フローチャートには詳細に記載されていないが、ステップS160で肯定判別された場合でも、ステップS150の領域抽出処理ルーチンの実行後でない場合には、ステップS180に処理を進める。ステップS180では、ツールバーBR2上の「保存」のボタンBT4が、マウス20によってクリック操作されたか否かを判別して、ここで、肯定判別されたときには、CPU30は、文字列認識処理ルーチンで得られたテキストデータを保存する（ステップS190）。ステップS190の実行後、またはステップS190で否定判別された場合には、「リターン」に抜けてこの制御処理のルーチンを一旦終了する。

【0051】図11は、ステップS170で実行される認識処理ルーチンの詳細を示すフローチャートである。図示するように、この認識処理ルーチンに処理が移行すると、CPU30は、まず、RAM32に格納された抽出画像データPDを文字列単位に分割する処理を行なう（ステップS300）。抽出画像データPDに示される文字列は、ウィンドウW1に設けられたラジオボタンRBT1、RBT2で横書きか縦書きかが指定されていることから、この指定の結果に応じた方向に抽出画像データPDを分割する。図12は、この抽出画像データPDを文字列単位に分割する様子を示す説明図である。図8に示したように「横書き」のラジオボタンRBT1が選択されている場合、図12に示すように、抽出画像データPDを縦方向に分割する。その分割する境BDは、連続する空白部分BKを選択して、その空白部分BKによって決まる。この結果、文字列単位に分けられた抽出画像データPD（以下、文字列抽出画像データPDLと呼ぶ）が得られる。

【0052】図11に戻り、ステップS300の実行後、CPU30は、変数gに初期値1をセットする（ステップS310）。次いで、CPU30は、上記変数gに対応する順番目（以下、第g行目と呼ぶ）の文字列抽出画像データPDLに対して文字認識を施す第1認識処理（後述する）を行なう（ステップS320）。この第1認識処理で文字認識に成功したか否かを判別して（ステップS330）、成功した場合にはステップS370に処理を進める。一方、第1認識処理で文字認識に失敗したと判定された場合には、CPU30は、第g行目の

文字列抽出画像データPDLに対して文字認識を施す第2認識処理(後述する)を実行する(ステップS340)。この第2認識処理で文字認識に成功したか否かを判別して(ステップS350)、成功した場合にはステップS370に処理を進める。一方、第2認識処理でも文字認識に失敗したと判定された場合には、CPU30は、第g行目の認識結果はエラーである旨を記憶する(ステップS360)。ステップS360の実行後、ステップS370に処理を進める。

【0053】ステップS370では、CPU30は、変数gを値1だけインクリメントする処理を行なう。その後、変数gが、ステップS300の分割の結果得られた行数LNを越えたか否かを判別する(ステップS380)。ここで、否定判別された場合には、ステップS320に処理を戻して、ステップS370でインクリメントされたgの値の行に属する分割データについて、ステップS320ないしS360の処理を繰り返して実行する。

【0054】ステップS380で、変数gがLNを越えたと判定された場合には、CPU30は、全ての行についての文字認識が終了したとして、全ての行についての文字認識の結果得られたテキストデータをワープロソフトウェア70に出力する(ステップS390)。その後、「リターン」に抜けてこの認識処理ルーチンの処理を一旦終了する。

【0055】図13は、ステップS320で実行される第1認識処理の詳細を示すフローチャートである。処理が開始されると、CPU30は、まず、ステップS300で文字列単位に分割された文字列抽出画像データPDLのうちの第g行目に属する文字列抽出画像データPDLを文字単位に分割する処理を行なう(ステップS400)。文字列抽出画像データPDLは、複数の文字が間隔を空けて配列された文字列の画像データであることから、文字列抽出画像データPDLから連続する空白部分を選択して、その空白部分によって文字を区分けすることによって文字単位に分割を行なう。この結果、文字単位に分けられた抽出画像データ(以下、文字抽出画像データPDCと呼ぶ)が得られ、これら文字抽出画像データPDCは、格納場所であるRAM32から順に読み出される。

【0056】次いで、CPU30は、第g行目に属する文字抽出画像データPDCを1つずつ順に、候補文字列辞書ファイル内の全ての文字データ(テキストデータ)により特定される標準文字パターンとそれぞれ比較する処理を行なう(ステップS405)。ここでいう候補文字列辞書ファイルは、HDD42に格納されている「都道府県」と「繊維(東京第1部)」の候補文字列辞書のファイルXD a、XD bのうちのステップS220で指定された辞書についてのものである。

【0057】図14は、ステップS405での比較処理

の様子を示す説明図である。図示するように、例えば、第g行目に属する文字抽出画像データPDCの配列が「東洋紡」の文字を読み取ったスキャナ画像であった場合、まず、第1番目である「東」といった文字抽出画像データPDCと、ステップS220で指定された辞書である「繊維(東京第1部)」の候補文字列辞書ファイルXD bとの間の比較の処理が行なわれる。この比較の処理は、詳細には、次のとおりに行なわれる。候補文字列辞書のファイルXD b中の第1番目の候補文字列データの第1の文字データ(「片」を示すテキストデータ)と対応する標準文字パターンp m1を標準文字パターンファイルPTFから抽出して、上記「東」といった文字抽出画像データPDCとこの抽出した標準文字パターンp m1とのパターン照合が第1番目に行なわれる(図中㊶)。次いで、候補文字列辞書ファイルXD b中の第1番目の候補文字列データの第2の文字パターン(「倉」を示すテキストデータ)と対応する標準文字パターンp m2を標準文字パターンファイルPTFから抽出して、上記「東」といった文字抽出画像データPDCとこの抽出した標準文字パターンp m2とのパターン照合が行なわれる(図中㊷)。

【0058】その後、同様にして、候補文字列辞書ファイルXD b中の第2番目の候補文字列データについても一の文字データ毎に対応する標準文字パターンを抽出して、上記「東」といった文字抽出画像データPDCとこの抽出した標準文字パターンとパターン照合が行なわれる。このように順に、候補文字列辞書ファイルXD b中の全ての候補文字列データの全ての文字データの対応する標準文字パターンとの間で上記「東」といった文字抽出画像データPDCのパターン照合が行なわれる。このパターン照合によって各標準文字パターンとの間の類似度が求められ、照合の結果として、類似度が所定値以上となった標準文字パターンについての文字データが、類似度が高いものから順に検出されることになる。

【0059】次いで、第g行目に属する文字抽出画像データPDCの配列のうちの第2番目の文字抽出画像データPDC(図14では「洋」と、ステップS220で指定された辞書である「繊維(東京第1部)」の候補文字列辞書ファイルXD bとの間の比較の処理が行なわれる。詳細には、上記第1番目の文字抽出画像データPDCに対する比較の処理と同様に、候補文字列辞書ファイルXD b中の全ての候補文字列データの全ての文字データに対応する標準文字パターンとの間でパターン照合が行なわれる。このようにして、第g行目に属する全ての文字抽出画像データPDCについて、候補文字列辞書ファイルXD b中の全ての候補文字列データの全ての文字データに対応する標準文字パターンとの間でそれぞれパターン照合が行なわれる。

【0060】ステップS405の結果、第g行目に属する各文字抽出画像データPDC毎に、類似度の高い文字

データが一または複数検出されることになる。ステップS405の実行後、CPU30は、まず、全ての文字抽出画像データPDCについての第1番目に類似度が高い文字データを集めて、最も類似性の高い比較用文字列データCDLを作成する(ステップS410)。次いで、CPU30は、その比較用文字列データCDLと、ステップS220で指定された辞書である「織維(東京第1部)」の候補文字列辞書ファイルXDb中の全ての候補文字列データとを順に比較する処理を行なう(ステップS420)。

【0061】ステップS420の比較処理により、その比較用文字列データCDLと一致する、候補文字列辞書ファイルXDb中の候補文字列データが見つかった場合(ステップS430)、CPU30は、その一致する候補文字列データを認識結果としてRAM32に記憶する(ステップS440)。その後、「リターン」に抜けて、この処理のルーチンを一旦終了する。一方、ステップS430で一致するものがないと判別された場合には、ステップS450に処理を進める。

【0062】ステップS450では、ステップS410で作成した比較用文字列データCDLについて一文字だけ類似度の高い次の文字データに変更して、新たな比較用文字列データCDLを作成する。例えば、ステップS410で作成した比較用文字列データCDLが「菱洋紡」であった場合に、第1文字目について、第1番目に類似度の高い文字データである「菱」から第2番目に類似度の高い文字データである「東」に変更して、「東洋紡」といった新たな比較用文字列データCDLを作成する。

【0063】その後、CPU30は、ステップS450で新たな比較用文字列データCDLが作成されたか否かを判定し(ステップS460)、ここで、作成されたと判定されたときには、ステップS420に処理を戻して、その新たな比較用文字列データCDLについての比較の処理を行なう。ステップS430でこの比較処理においても一致するとの判定が得られなかったときは、再度、ステップS450で、比較用文字列データCDLについて一文字だけ次の類似度の高い文字データに変更して新たな比較用文字列データCDLを作成し、その後、ステップS420に処理を戻す。先にこのステップS450の処理を実行したときには、第1文字目について第2番目に類似度の高い文字データに変更する処理を行なったが、次の実行時のステップS450では、第2文字目について第2番目に類似度の高い文字データに変更する処理を行なう。ステップS450では、このように1文字ずつ文字データを変更して新たな比較用文字列データCDLを作成する。

【0064】ステップS450で、比較用文字列データCDLを構成する各文字がステップS405で検出された全ての類似度が高い文字データに置き換えられると、

その後、ステップS460で新しい組合せの比較用文字列データCDLが作成されなかったと判断される。この場合には、CPU30は、ステップS470に処理を進めて、認識が不可能であった旨を認識結果としてRAM32に記憶する。その後、「リターン」に抜けて、この処理のルーチンを一旦終了する。

【0065】図15および図16は、ステップS340で実行される第2認識処理の詳細を示すフローチャートである。図15に示すように、CPU30は、処理が開始されると、まず、変数mを値1にセットする(ステップS500)。次いで、候補文字列辞書ファイルに格納される候補文字列データの数を変数Pにセットする(ステップS510)。ここでいう候補文字列辞書ファイルは、HDD42に格納されている「都道府県」と「織維(東京第1部)」の候補文字列辞書のファイルXDa、XDbのうちのステップS220で指定された辞書についてのものである。

【0066】続いて、CPU30は、上記候補文字列辞書ファイルの中から変数mの値に対応する順番目の候補文字列データの文字データの数(文字数)を数えて、この文字数を変数Sにセットする(ステップS520)。この実施例では、候補文字列辞書ファイルに格納される各候補文字列データの文字数を上述したように数える構成としたが、これに替えて、候補文字列辞書のファイルを、候補文字列データと当該候補文字列データの文字数とが予め対に記憶されたデータ構造として、その文字数を変数Sにセットする構成とすることもできる。

【0067】ステップS520の実行後、CPU30は、ステップS300で文字列単位に分割された文字列抽出画像データPDLのうちの第g行目に属する文字列抽出画像データPDLを上記変数Sの数に等分割する処理を行なう(ステップS530)。図17および図18は、文字列抽出画像データPDLを変数Sの数に等分割する様子をそれぞれ示す説明図である。図17に示すように、例えば、文字列抽出画像データPDLが「東洋紡」の文字を読み取ったスキャナ画像であり、候補文字列辞書ファイルXDbのm番目の候補文字列データが「片倉」であった場合、文字列抽出画像データPDLをその文字列の並び方向(横方向)に、「片倉」の文字数である値2の数に均等に分割する。詳細には、文字列抽出画像データPDLの横方向の長さxを求めて、一つの部分の横方向がxを値2で割った大きさとなるように均等に分割する。この結果、第1番目の文字抽出画像データPDC(1)と第2番目の文字抽出画像データPDC(2)が得られる。

【0068】図18に示すように、例えば、文字列抽出画像データPDLが「東洋紡」の文字を読み取ったスキャナ画像であり、候補文字列辞書ファイルXDbのm番目の候補文字列データが「グンゼ」であった場合、文字列抽出画像データPDLをその文字列の並び方向に、

「グンゼ」の文字数である値3の数に均等に分割する。この結果、第1番目ないし第3番目の文字抽出画像データPDC(1), PDC(2), PDC(3)が得られる。これらの例示のように、ステップS530の結果、第g行目に属する文字列抽出画像データPDLは、候補文字列辞書ファイルXD_bのm番目の候補文字列データの文字数Sの数に等分割されて、特許請求の範囲で言うところの部分画像データに相当する文字抽出画像データPDC(1)…PDC(S)が得られる(Sが1の場合、PDC(1)だけとなる)。これら文字抽出画像データPDC(1)…PDC(S)は、格納場所であるRAM32から個別に読み出される。

【0069】ステップS530の実行後、CPU30は、変数nを値1にセットして(ステップS540)、その後、ステップS530で得られた文字抽出画像データPDCの内の変数nの値に対応する順番目の候補文字列データPDC(n)を、候補文字列辞書ファイルと比較する処理を行なう(ステップS550)。詳細には、ステップS220で指定された候補文字列辞書ファイルの中の変数mの値に対応する順番目の候補文字列データ(XD_a(m)またはXD_b(m))の内の変数nに対応する順番目の文字データと対応する標準文字パターンp_tを標準文字パターンファイルPTFから抽出して、この抽出した標準文字パターンp_tと上記文字抽出画像データPDC(n)とのパターン照合を行なう。このパターン照合によって、文字抽出画像データPDC(n)とその標準文字パターンp_tとの間の類似度が求められ、その求められた類似度はRAM32に記憶される。

【0070】その後、CPU30は、変数nを値1だけインクリメントして(ステップS560)、その変数nが上記変数Sを越えたか否かを判別する(ステップS570)。ここで、否定判別された場合には、ステップS550に処理を戻して、ステップS560でインクリメント後の変数nに従う文字抽出画像データPDC(n)とm番目の候補文字列データの内のn番目の文字データの標準文字パターンとの照合を行なう。ステップS570で肯定判別、すなわち、変数nが候補文字列データの文字数である変数Sの値を越えた場合には、第g行目に属する文字列抽出画像データPDLから分割された全ての文字抽出画像データPD(1)…PDC(S)についてのパターン照合が終了したとして、ステップS580に処理を進める。

【0071】ステップS580では、ステップS550で求めた、文字抽出画像データPDC(1)…PDC(S)についての類似度を合計して、その合計値を変数Sの値で割った値を、m番目の候補文字列データに対する類似度SM(m)として記憶する。その後、CPU30は、変数mを値1だけインクリメントして(ステップS590)、その変数mが上記変数Pを越えたか否かを

判別する(ステップS600)。ここで、否定判別された場合には、ステップS510に処理を戻して、ステップS590でインクリメント後の変数mに従うステップS510ないしS580の処理を繰り返し実行する。ステップS600で肯定判別、すなわち、変数mが候補文字列辞書ファイル文字列データの数である変数Pの値を越えた場合には、候補文字列辞書ファイルに格納された全ての候補文字列データとのパターン照合が終了したとして、図16のステップS610に処理を進める。

【0072】ステップS610では、CPU30は、ステップS580で算出された各候補文字列データに対する類似度SM(1)～SM(P)の中から最大のものを選択する。次いで、CPU30は、その最大の類似度SMが予め定めた閾値SM_bを上回っているか否かを判別して(ステップS620)、ここで、上回っているかと判別された場合には、その最大の類似度SMとなった候補文字列データを認識結果としてRAM32に記憶する(ステップS630)。その後、「リターン」に抜けて、この処理のルーチンを一旦終了する。一方、ステップS620で最大の類似度SMが閾値SM_b以下であると判別された場合には、CPU30は、ステップS640に処理を進めて、認識が不可能であった旨を認識結果としてRAM32に記憶する。その後、「リターン」に抜けて、この処理のルーチンを一旦終了する。

【0073】すなわち、上記構成の第2認識処理のルーチンによれば、まず、第g行目に属する文字列抽出画像データPDLは、候補文字列辞書ファイルに格納される各候補文字列データの文字数Sと一致する数に等分割される(ステップS520、S530)。その後、その分割によって得られた文字抽出画像データPDC(1)～PDC(S)は、候補文字列辞書ファイルに格納される各候補文字列データの同じ配列順位の文字データと、すなわち、図17、図18に示すように、PDC(1)は1番目の文字データと、PDC(2)は2番目の文字データと、…、PDC(S)はS番目の文字データとそれぞれ標準文字パターンp_tに基づくパターン照合がなされる(ステップS540～S570)。そうしてそのパターン照合の結果、最も類似度が高いと判定された候補文字列データが、第g行目に属する文字列抽出画像データPDLの文字認識結果として記憶される。この文字認識結果である候補文字列データは、図10に示したステップS390でワープロソフトウェア70に出力される。

【0074】CPU30とCPU30で実行される上記ステップS520、S530の処理とが、この発明の部分画像データ切り出し手段に相当する。CPU30とこのCPU30で実行される上記ステップS540～S570の処理とが、この発明の照合手段に相当する。CPU30とこのCPU30で実行される上記ステップS580～S630の処理とが、この発明の出力制御手段に

相当する。

【0075】以上のように構成されたこの実施例では、文字列抽出画像データPDLから切り出した文字抽出画像データPDCの文字列抽出画像データPDL全体に対する占有割合を、候補文字列データにおける1文字の占める割合と一致させて部分画像データの切り出しを行なうことができる。このために、画像が記録された原稿PPに汚れ等があっても隣り合う2文字が1文字に相当すると誤判定することもない。したがって、文字認識の精度が極端に低下することを防止することができる。

【0076】また、この実施例では、前述したように、HDD42に、「都道府県」の候補文字列辞書のファイルXD_aと「繊維(東京第1部)」の候補文字列辞書のファイルXD_bとの2つの辞書が予め格納されており、ディスプレイ画面に表示されるウィンドウW1の入力欄I Aからいずれかの辞書を選択可能となっている。このために、適切な辞書を選択することで、いずれの範疇の文字列をも高精度に認識することができる。

【0077】さらに、この実施例では、前述したように、第1認識処理により、文字列抽出画像データPDLの中から空白部分に基づいて切り出した文字抽出画像データPDC毎に、候補文字列データとのパターン照合(詳細には、候補文字列データによって特定される各標準文字パターンとのパターン照合)を行ない、この照合結果が認識不可となった場合に、第2認識処理により、文字列抽出画像データPDLの中から、候補文字列データの文字数Sと一致する数に等分割した文字抽出画像データPDC毎に、候補文字列データとの照合(詳細には、候補文字列データによって特定される各標準文字パターンとのパターン照合)を行なう。したがって、この実施例では、文字列抽出画像データPDLの中から空白部分に基づいて部分画像データを切り出す一般的な切り出し方法に基づく文字認識と、本発明に関わる文字認識との2段階をもって文字認識を行なうことから、認識精度を一層高めることができる。

【0078】本発明の他の実施形態について、次に説明する。上記実施例では、第1認識処理と第2認識処理とで文字認識を行なうアルゴリズムが相違する構成としたが、これに替えて、第1認識処理と第2認識処理とのアルゴリズムは同一のものとすることもできる。すなわち、第1認識処理において、ステップS405～S470の処理を、第2認識処理のステップS500～S520、S540～S640にそっくり入れ替える構成とすることもできる。あるいは、第2認識処理において、ステップS510、S540～S580を削除して、ステップS610～S640の処理を、第1認識処理のステップS405～S470にそっくり入れ替える構成とすることもできる。

【0079】また、上記実施例では、候補文字列データとして、都道府県名、証券銘柄等を例示していたが、必

ずしもこういった単語を単位とする必要もなく、単語の集まりである文、文の集まりである段落等を表わす文字列を候補文字列データとすることもできる。この構成によれば、文単位、段落単位で文字列の認識が可能となる。

【0080】上記実施例では、スキャナ画像で示される記録紙上の文字は、文字幅が等しいフォントで記録されたものとしていた。これに替えて、記録紙上の文字がプロポーショナルな文字である記録されている場合もある。文字認識の精度が上がらない場合には、ステップS530において、等分割された文字抽出画像データPDC(1)…PDC(S)の横幅を微調整する処理を行なう。すなわち、ステップS530で等分割した分割の位置を、図19に示すように、等分割の境界位置LXから文字列の並び方向に微少に移動した位置LX'に変える。そうして、得られた新たな文字抽出画像データPDC'(1)…PDC'(S)を用いて、再度ステップS540ないしステップS640の処理が行なわれる構成とする。このように等分割された文字抽出画像データPDC(1)…PDC(S)の横幅を少しずつ微調整しながらパターン照合および出力制御を繰り返す行なう。この構成によって、画像データにて示される文字のフォントがプロポーショナルな文字である場合にも、高精度に文字を認識することができる。

【0081】上記実施例では、この発明の文字列認識装置として、パーソナルコンピュータを用いていたが、これに替えて、スキャナ装置にこのパーソナルコンピュータの機能を持たせた構成とすることもできる。また、上記実施例では、光学的に読み取って得られた画像データは、イメージスキャナ14から入力する構成としていたが、これに替えて、ネットワークを介して外部から取り込む構成とすることもできる。

【0082】以上、本発明の一実施例を詳述してきたが、本発明は、こうした実施例に何等限定されるものではなく、本発明の要旨を逸脱しない範囲において種々なる態様に実施することができるのは勿論のことである。

【図面の簡単な説明】

【図1】この発明の一実施例を適用するコンピュータシステムのハードウェアの概略構成を示すブロック図である。

【図2】コンピュータ本体16によって実行される文字認識処理の様子を示すブロック図である。

【図3】文字認識用ソフトウェア50に従う制御処理を示すフローチャートである。

【図4】文字認識用ソフトウェア50の起動後の初期画面のウィンドウW0を示す説明図である。

【図5】原稿PPの一例を示す説明図である。

【図6】ステップS150で実行される領域抽出処理ルーチンの詳細を示すフローチャートである。

【図7】原稿PPのスキャン画像の中から、東京第1部の「繊維」の銘柄を表わす範囲の画像領域PAを指定する操作を示す説明図である。

【図8】抽出画像データPDが表示されるウィンドウW1の一例を示す説明図である。

【図9】HDD42およびRAM32に格納されるこの発明に関わるデータファイルを示す説明図である。

【図10】CRTディスプレイ12に表示される追加用のウィンドウW2を示す説明図である。

【図11】ステップS170で実行される認識処理ルーチンの詳細を示すフローチャートである。

【図12】抽出画像データPDを文字列単位に分割する様子を示す説明図である。

【図13】ステップS320で実行される第1認識処理の詳細を示すフローチャートである。

【図14】ステップS405で実行される比較処理の様子を示す説明図である。

【図15】ステップS340で実行される第2認識処理の前半部分を示すフローチャートである。

【図16】前記第2認識処理の後半部分を示すフローチャートである。

【図17】文字列抽出画像データPDLを値2である変数Sの数に等分割する様子を示す説明図である。

【図18】文字列抽出画像データPDLを値3である変数Sの数に等分割する様子を示す説明図である。

【図19】文字抽出画像データPDC(1)…PDC(3)の横幅を微調整する例を示す説明図である。

【符号の説明】

12…CRTディスプレイ

14…イメージスキャナ

16…コンピュータ本体

18…キーボード

20…マウス

31…ROM

32…RAM

33…表示画像メモリ

34…マウスインタフェース

35…キーボードインタフェース

36…CDC

37…HDC

38…CR TC

40…入出力機用インタフェース

41…I/Oポート

42…ハードディスクドライブ

44…モデム

46…公衆電話回線

47…サーバ

48…候補文字列記憶部

49…標準文字パターン記憶部

50…文字認識用ソフトウェア

51…スキャン画像取込部

52…領域抽出部

53…候補文字列指定部

54…候補文字列記憶部

55…第1認識部

56…文字列出力部

57…第2認識部

60…スキャナドライバ

70…ワープロソフトウェア

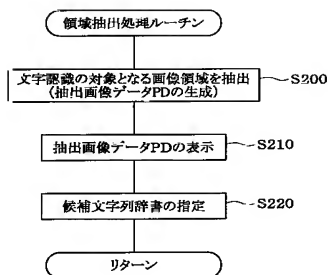
80…ディスプレイドライバ

XDa、XD b…候補文字列辞書ファイル

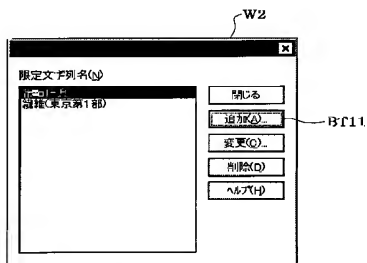
PDL…文字列抽出画像データ

PDC…文字抽出画像データ

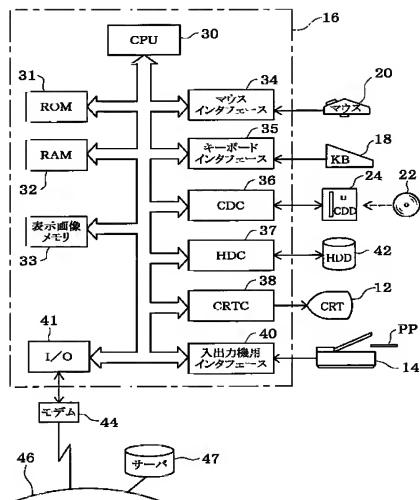
【図6】



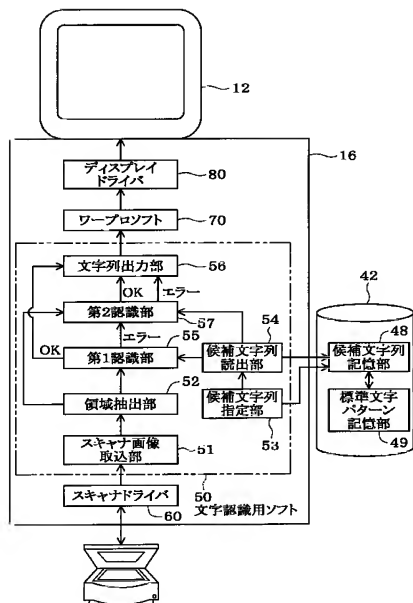
【図10】



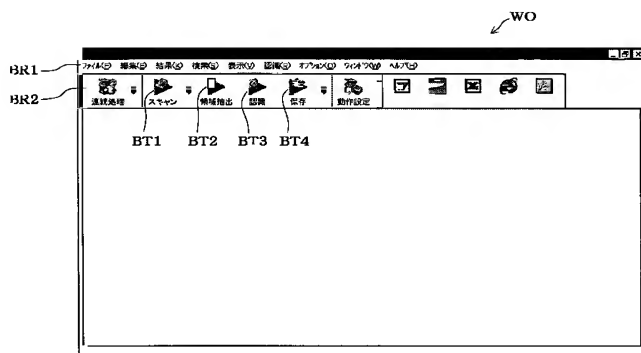
【図1】



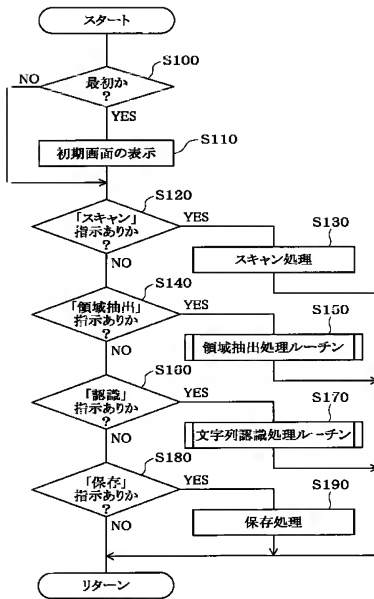
【図2】



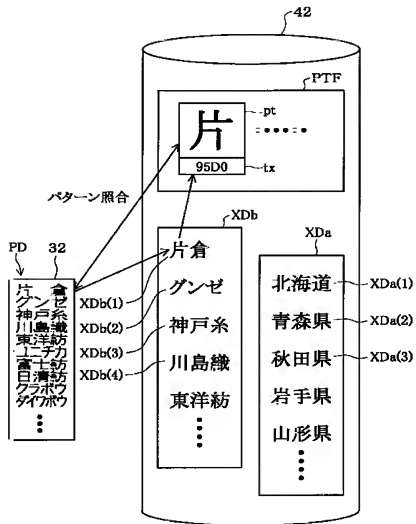
【図4】



【図3】

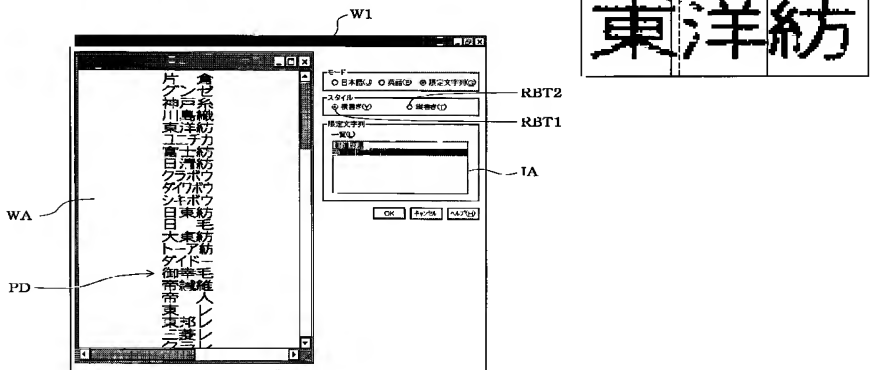


【図9】



【図19】

【図8】



【图5】

PP

23 証券1 11版

【第三种郵便物認可】

東京第1部

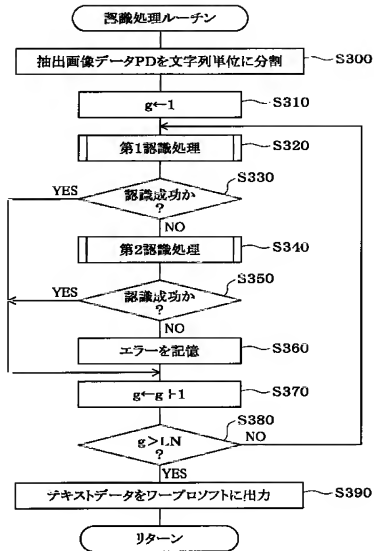
2月5日
(月曜日)

[illegible]

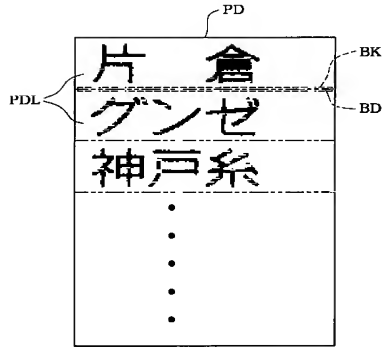
【図7】

業		4340A100: 65.9	A加ド青	2855	2855	2820	2830	110	24.7	ドワベ
87	▲ 1	348	・ヨコレイ	630	630	628	628	▲ 9	12	・中国塗
67	0	37	・東洋水	969	969	920	920	▲ 50	48	・日特塗
191	▲ 1	1791	A日清食	2590	2630	2580	2600	▲ 10	81.1	・AA種イキ
490	▲ 1	270	・フジ	750	750	740	797	▲ 47	26	・インキ
1106	▲ 47	1425.0	ARハイパ	1295	1295	1295	1296	▲ 34	2.9	・カ列IX
486	▲ 9	34	・ミヨシ	4700	4710	4510	4610	▲ 42	1100	・洋イネ
117	0	40		8070	8250	8070	8130	▲ 40	32	・富士号
業		4340A100: 65.9	A加ド青	2855	2855	2820	2830	110	24.7	ドワベ
341	▲ 6	1.0	・高橋	466	466	463	463	▲ 5	5	・養生堂
356	▲ 6	7.7	・神戶糸	410	410	403	407	▲ 4	219	・ライオン
105	▲ 8	2.0	・川島織	125	128	125	125	0	94	・高砂香
184	▲ 1	3	・東洋紡	225	226	219	224	▲ 2	30	・AFアガル
209	▲ 3	89	・ユニテカ	82	82	80	82	▲ 2	1158	・コニセー
412	▲ 4	2121	・富士紡	52	52	51	51	▲ 1	63	・エステル
298	▲ 4	638	・日清紡	535	536	521	523	▲ 7	181	・Aコニシ
46	▲ 2	630	・クワボフ	172	172	174	174	▲ 3	498	・小林製薬
33	▲ 1	144	・シキボフ	66	67	64	65	▲ 2	107	・イラクミ
33	▲ 1	464	・日東紡	142	143	140	141	▲ 2	226	・北興化
52	▲ 0	1329	・日東毛	411	412	405	405	▲ 6	41	・クアイル
274	▲ 1	12	・大東紡	75	75	73	73	▲ 2	81	・アロ化成
105	▲ 1	851	・タイト	245	245	239	245	▲ 3	81	・アロ化成
38	▲ 0	124	・帝華毛	400	400	392	395	▲ 3	4	・文部科学
93	▲ 2	77	・帝華織	166	166	166	166	▲ 5	3	・アコ
184	▲ 1	10	・帝華人	511	524	510	519	▲ 1	974	・バルカム
116	▲ 1	4	・東邦レ	414	416	409	411	▲ 3	1140	・Aニテム
125	▲ 2	4	・三ツレ	387	387	381	382	▲ 6	76	・日三
357	▲ 1	413	・カイトベ	930	930	912	913	▲ 18	533	・昭和石
47	▲ 2	131	・住江織	127	136	127	135	▲ 6	141	・興亜石
140	▲ 2	32	・アールト	350	351	350	351	▲ 4	1	・コス石
59	▲ 2	100	・市川毛	243	243	243	243	▲ 3	5	・富士興
390	▲ 0	62	・バリス	222	222	216	216	▲ 6	12	・ニルキ
111	▲ 1	35	・日東網	106	113	106	111	▲ 1	65	・東亜セ
67	▲ 0	14	・日東工	66	67	66	67	▲ 1	9	・Jエナ
365	▲ 3	256	・アツキ	178	180	177	177	▲ 5	13	・浜ゴム
207	▲ 0	2.0	・アツキ	71	72	70	71	▲ 1	68	・洋ゴム
193	▲ 0	9	・アツキ	152	153	151	151	▲ 0	46	・オーツ
136	▲ 1	8	・アツキ	302	302	300	300	▲ 5	4	・カサド
75	▲ 5	10	・東海染	92	92	91	91	▲ 1	77	・住友ソ
127	▲ 0	10	・小松精	273	273	268	268	▲ 3	77	・オカト
90	▲ 1	15	・コル	953	977	953	971	▲ 18	115	・ニッタ
195	▲ 0	15	・キメテ	6200	6370	6200	6360	▲ 160	47.7	・東海ゴ
128	▲ 5	17.6	・ワールド	4440	4600	4400	4560	▲ 30	2.6	・ニ星
91	▲ 1	160	・ワールド	1051	1053	1036	1051	▲ 9	129	・バド
470	▲ 18	95	・ワールド	73	74	73	74	▲ 0	15	・旭硝子
32	▲ 1	1274	・ワールド	910	923	901	915	▲ 10	169	・板硝子
219	▲ 3	74	・市田	60	65	60	64	▲ 4	1510	・石塚硝
135	▲ 1	3	・市田	98	99	95	96	▲ 2	417	・山時硝
219	▲ 3	50	・市田	385	385	385	385	▲ 1	2	・ニ星
180	▲ 2	8	・市田	116	116	116	116	▲ 2	43	・ニ星
400	▲ 2	8	・市田	116	116	116	116	▲ 2	8	・日硝硝
350	▲ 10	7	・市田	139	141	116	1123	▲ 12	86	・住友大
354	▲ 15	1.9	・市田	184	185	177	184	▲ 0	73	・太平洋
2280	▲ 45	344.4	・市田	166	161	155	161	▲ 3	39	・第一セ
業		4340A100: 65.9	A加ド青	2855	2855	2820	2830	110	24.7	ドワベ
490	▲ 13	58	・東海パ	381	382	380	380	0	14	・日本ビ
185	▲ 1	68	・王子紙	538	540	530	534	▲ 4	1317	・日コ
469	▲ 2	36	・日本紙	596	601	585	585	▲ 16	1812	・東海カ
126	▲ 3	99	・三豊紙	225	228	223	224	▲ 1	178	・カーボ
29	▲ 1	404								・リタケ

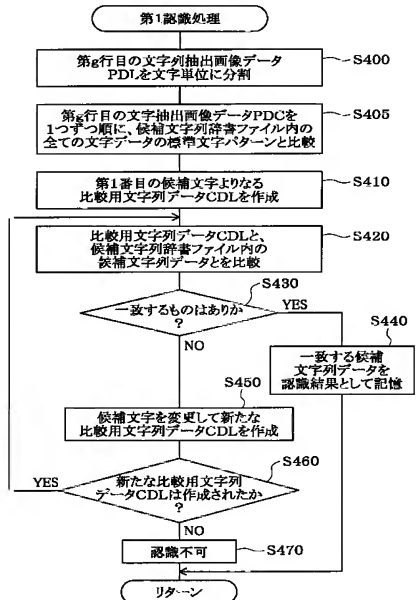
【図11】



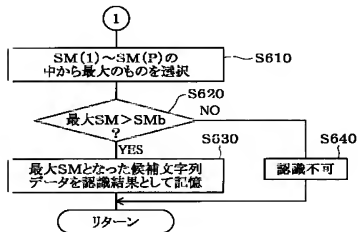
【図12】



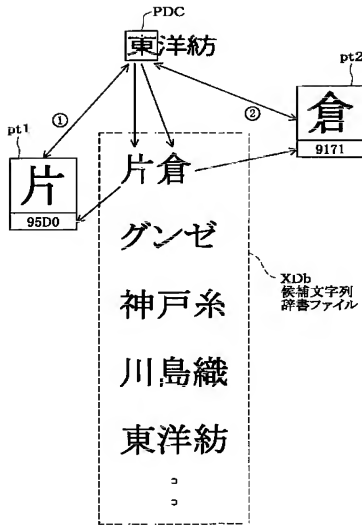
【図13】



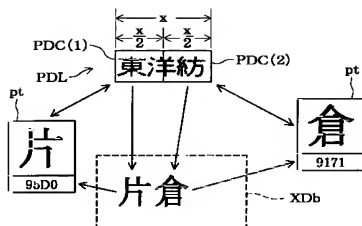
【図16】



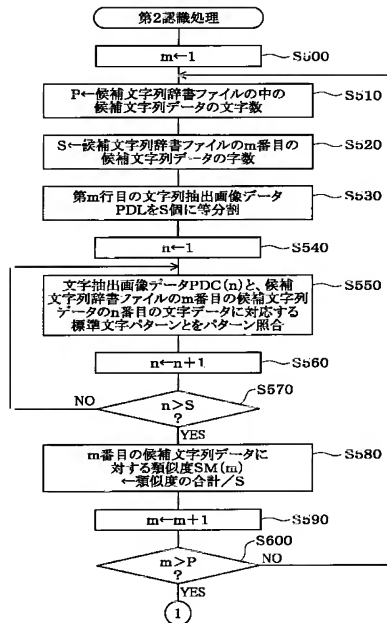
【図14】



【図17】



【図15】



【図18】

